PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 12/00, 13/00 — A1

(11) International Publication Number: **WO 97/09676**

(43) International Publication Date: 13 March 1997 (13.03.97)

(21) International Application Number: PCT/US96/13781

(22) International Filing Date: 29 August 1996 (29.08.96)

(30) Priority Data:
08/522,903 1 September 1995 (01.09.95) US

(71) Applicant: EMC CORPORATION [US/US]; 171 South Street, Hopkinton, MA 01748-9103 (US).

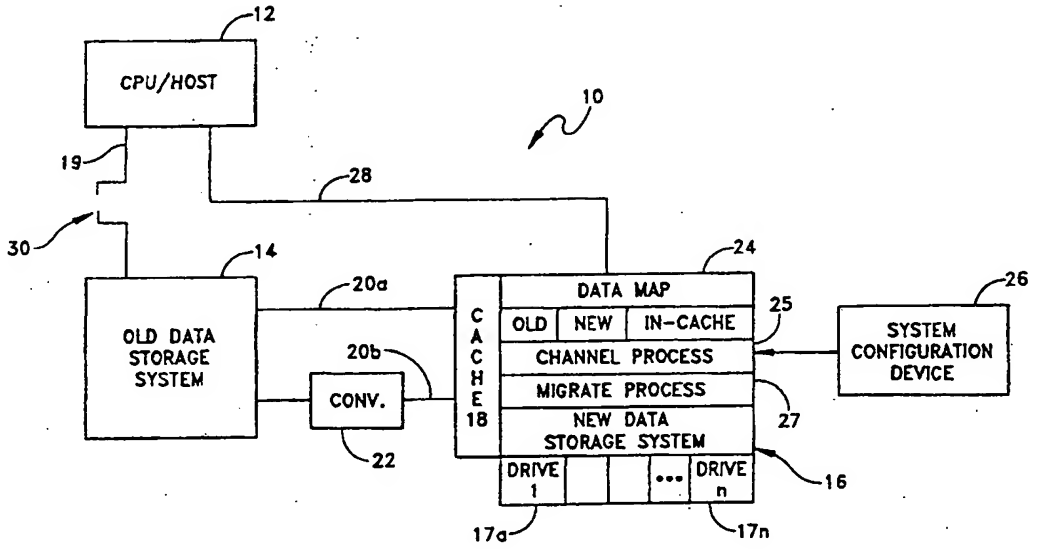
(72) Inventors: OFEK, Yuval; 13 Forest Lane, Hopkinton, MA 01748 (US). YANAI, Moshe; 15 Catlin Road, Brookline, MA 02146 (US).

(74) Agent: HERBSTER, George, A.; Pearson & Pearson, 12 Hurd Street, Lowell, MA 01852 (US).

(81) Designated States: JP, KR, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: SYSTEM AND METHOD FOR ON-LINE, REAL-TIME, DATA MIGRATION



(57) Abstract

A system and method (25, 27) for providing on-line, real-time, transparent data migration from a first data storage system (14) to a second data storage system (16). The second data storage system is interposed between a host (12) and the first data storage system. A data map (24) identifies data elements stored on the second data storage system and corresponding data elements copied from the first to the second data storage system. In response to a host data request, the second data storage retrieves the data if stored therein. Otherwise, the second data storage system retrieves the data from the first data storage system, writes the data to itself and updates the data map. When not busy servicing requests, the second data storage system copies data from the first to the second data storage device independently of any coupled host.

Background of the Invention

System and Method For On-Line, Real Time, Data Migration

Technical Field

This invention relates to data storage systems and more particularly, to a system and method for on-line replacement of an existing data storage subsystem.

Background Art

Data processing centers of businesses and organizations such as banks, airlines and insurance companies, for example, rely almost exclusively on their ability to access and process large amounts of data stored on a data storage device. Data and other information which is typically stored on one or more data storage devices which form part of a larger data storage system is commonly referred to as a database.

Databases are nearly always "open" and constantly "in use" and being accessed by a coupled data processing system, central processing unit (CPU) or host mainframe computer. The inability to access data is disastrous if not a crisis for such business and organizations and will typically result in the business or organization being forced to temporarily cease operation.

During the course of normal operations, these businesses and organizations must upgrade their data storage devices and data storage systems. Although such upgrading sometimes includes only the addition of data storage capacity to their existing physical systems, more often than not upgrading requires the addition of a completely separate and new data storage system. In such cases, the existing data on the existing data storage system or device must be backed up on a separate device such as a tape drive, the new system installed and connected to the data processing unit, and the data copied from the back-up device to the new data storage system. Such activity typically takes at least two days to accomplish. If the conversion takes more than two days or if the business or organization cannot withstand two days of inoperability, the need and desire to upgrade their data storage system may oppose an insurmountable problem.

first data storage device which was previously coupled to an external source of data including a data processing device such as a host computer, or a network which may be connected to a number of data processing devices such as a number of host computers. The data processing device such as a host computer reads data from and writes data to the data storage device. The first data storage device initially includes a plurality of data elements currently being accessed by the data processing device.

At least one second data storage device is provided which is coupled to the first data storage device and to the data processing device, for storing data elements to be accessed by the data processing device. The second data storage device preferably includes a data element map including at least an indication of whether or not a particular data element is stored on the second data storage system.

In one embodiment, the second data storage system independently migrates data from the first to the second data storage system independent of the source. In another embodiment, the second data storage system is responsive to the external source, for migrating data from the first to the second data storage system.

In yet another embodiment, the data processing device issues a data read request (in the case of a read data operation), or a data write command (in the case of a write operation). The request is received by the second data storage device. In the case of a read operation, second data storage device examines the data map or table to determine whether or not the data has been migrated to and is stored on the second data storage device. If it is determined that the data is stored on the second data storage device, the data is made available to the requesting device.

If the data is not stored on the second data storage device, the second data storage device issues a data request, in the form of a read data command, to the first data storage device, obtains the data and makes the data available to the requesting device. The data received from the first data

requesting device, and completely transparent to the coupled data processing device.

In the preferred embodiment, the second data storage device further includes or is coupled to a data storage device system configuration device, such as a computer, which provides configuration data to the data element map or table on the second data storage device, allowing the second data storage device to be at least partially configured in a manner which is generally similar or identical to the first data storage device.

Additionally, the preferred embodiment contemplates that the second and first data storage devices are coupled by a high speed communication link, such as a fiber optic link employing the "ESCON" communication protocol. The preferred embodiment also contemplates that the data storage device includes a plurality of data storage devices, such as disk drives. In this case, data elements may include one or more of a disk drive volume, track or record.

Brief Description of the Drawings

These and other features and advantages of the present invention will be better understood by reading the following detailed description, taken together with the drawings wherein:

Fig. 1 is a schematic diagram of an exemplary data processing and data storage system on which the system and method for providing on-line, data transparent data migration between first and second data storage systems in accordance with the present invention may be accomplished;

Fig. 2 is a schematic illustration of a data element map or table;

Fig. 3 is a flowchart outlining the steps of providing on-line, transparent data migration between first and second data storage systems according to the method of the present invention; and

Fig. 4 is a flowchart illustrating the steps for providing data migration between first and second data storage systems without data storage device or host system intervention when the second data storage device is not busy

data storage system 14, the use of such a data map/table will be explained in greater detail below.

The second data storage system 16 is typically and preferably coupled to a data storage system configuration device 26 such as a computer, which allows the user to configure the second data storage system 16 and the data map/table 24 as desired by the user. In the preferred embodiment, the second data storage system 16 is at least partially configured exactly as the first data storage system 14 is configured in terms of the number of logical devices, storage size, storage system type (3380/3390, for example) etc.

In the preferred embodiment, the data storage system configuration device 26 allows the user to configure at least a portion of the data storage area on second data storage system 16 to include data element storage locations or addresses which correspond to data element storage addresses on the first data storage system 14.

In the preferred embodiment, the second data storage system 16 is a disk drive data storage system employing a large number of fixed block architecture (FBA) formatted disk drives 17a-17n, and adapted for storing large amounts of data to be accessed by a host computer or other data processing device 12. The exemplary second data storage system 16 also typically includes a cache memory 18 which serves to hold or buffer data read and write requests between the second data storage system 16 and the host or other data processing device 12. Such data storage systems are well known to those skilled in the art and include, for example, the Symmetrix 5500 series data storage system available from EMC Corporation, Hopkinton, Massachusetts, a description of which is incorporated herein by reference.

Initially, the second or new data storage system 16 is first coupled to the first data storage system 14 by means of one or more data communication links or paths 20a, 20b. After the second data storage system 16 has been configured using a system configuration device 26 or other similar or equivalent device, or by the host 12, the second data storage

Such a hierarchical data map/table 24 is further explained and exemplified herein as well as in U.S. Patent Nos. 5,206,939 and 5,381,539 assigned to the assignee of the present invention and both fully incorporated herein by reference.

If the data is already stored in the second data storage system 16, the second data storage 16 retrieves the data (perhaps temporarily storing the data in cache memory 18) as is well known in the art, and makes the data available to the host or other requesting data processing device 12.

If the requested data is not on the second data storage system 16, channel or real-time data handling process 25 of the second data storage system 16 issues a read data request to the first data storage system 14 in the manner and format native or known to the first data storage system 14 (for example, standard IBM data read commands). Channel or real-time data handling process 25 is, in the preferred embodiment, a software program comprising a series of commands or instructions which receives one or more commands from the second data storage system interface to the host or CPU (typically called a "channel"), interprets those commands, and issues one or more corresponding commands which can be acted upon by the first data storage system. Such an 'interpreter' type of software is well known to those skilled in the art.

The first data storage system 14 then retrieves the requested data and provides it to the second data storage system 16. The second data storage system 16 then makes the data available to the host or other data processing unit 12 which has requested the data.

Since the second data storage system now has a copy of the data, the data will be written to the second data storage system 16 and the appropriate data map/table 24 flags or bits updated to indicate that the data has been migrated to the second data storage system 16 so that next time the same data element is requested, the second data storage system 16 will have the data already stored on the system and will not have to request it from the first data storage system.

hierarchical format in the data map/table 24. Thus, whenever the second data storage system 16 desires or needs to obtain information about a particular data element (be it an individual data record, track or volume), the data storage system 16 scans the data map/table 24 beginning at the device level 50 to determine whether or not the desired criterion or characteristic has been established for any track or volume of a device.

There will be a 'flag' or other similar indicator bit set, or other indication of the desired characteristic in the device entry 50, in the volume entry 52 and in the appropriate track entry 54 if the desired characteristic is found in that portion of the data storage device represented by the data map/table 24.

For example, the preferred embodiment of a data map/table 24 includes a write pending flag or bit 61 which is set if a particular data element is presently stored in cache 18 of the second data storage system 16 and must be written to longer term storage such as a disk drive 17a-17n. For exemplary purposes, assuming that track 2 of volume 1 is in cache 18 in the second data storage system 16 and write pending, the write pending flag or bit 61 and the in cache bit 58 at line entry 54b (for track two) will be set, as will the write pending bit 61 of volume 1 at line 52 of the data map/table 24, as will the write pending bit 61 of the device at line 50.

Thus, if the second data storage system 16 wishes to determine whether or not a particular track or record which has been requested is write-pending or has been migrated to the second system or of the status of some other attribute or characteristic, the data storage system 16 first determines which device or disk drive 17a-17n the data element is stored on and then checks the appropriate indicator flag bit for that device. If the particular indicator flag bit is not set for that device, then the second data storage system 16 knows immediately that no lower level storage unit or location such as a volume or track in that device has that attribute. If any lower data storage element in the hierarchical structure

next determines if the request or command is a read or a write request, step 101. If the command is a read command, the channel handling process 25 of the second data storage system 16 next determines if the requested data is already stored in the second data storage system 16, step 102, by reading its data table map/table 24.

If the data is stored on the second data storage system, step 102, the second data storage system 16 will make the data available to the host or other requesting data processing device 12, step 104, and return to step 100 to await receipt of a new data read or write request.

If, however, at step 102, the second data storage system 16 determines that the data is not presently stored on the second data storage system 16, the second data storage system 16 will generate a request to the first data storage system 14 to read the data, step 106.

The command or request to read data from the first data storage system 14 takes the same form as a read data command which would be issued from the host 12. Thus, for example, if the host 12 is an IBM or IBM compatible host or data processing device, the second data storage system 16 will issue an IBM compatible "read" command to the first data storage system 14. The channel and migrate processes 25,27 of the second data storage system 16 maintain a list of commands native to the first data storage system 14 and can easily convert command types, if necessary, from a first command type issued by the host 12 and understood by the second data processing system 16, to a second command type understood by the first data storage system 14.

Subsequently, the second data storage system 16 receives the requested data from the first data storage system 14, step 108 and writes the data to the cache memory 18 of the second data storage system 16 while updating the data element map/table 24, step 110. The second data storage system 16 then provides an indication to the host or data processing device 12 that the data is ready to be read, step 112. Subsequently, the second data storage system 16 will write the data from cache memory 18 to a more permanent storage location, such as a disk drive, on the second data storage

written is then written into the proper memory location in cache memory 18 (the occurrence of the actual "write" command), the data table/map 24 updated (for example, to indicate that the data is in cache memory 18 [data in cache bit set], that a write is pending on this data [write pending bit set], and that the data elements have been migrated [data needs migration bits re-set]) and the host or other central processing unit 12 informed that the write command is complete.

At some later time, the data in cache memory 18 which has been flagged as write pending is copied to a more permanent storage location, such as a disk drive, and the write pending bit reset.

Typically, data write requests are performed to update only a portion of the total or complete number of data elements stored in a predetermined data storage element or physical/logical confine (such as a disk drive track). The present invention, however, also realizes that in some cases, such as when the host or data processing unit 12 provides an indication that both the data structure (format) as well as the actual data contents are to be updated, reading old data from the first data storage system 14 may be eliminated since all data and data format or structure will be updated with the new write request. Such a data and format write command is so infrequent, however, that the preferred embodiment contemplates that each write request will cause a write request to be read from the first data storage system 14.

The method of present invention also allows the second or new data storage system 16 to provide transparent or "background" data migration between the first data storage system 14 and the second data storage system 16 irrespective of or in parallel with the data transfer or migration caused by the channel process which is serving the "channel" between the host 12 and the second data storage system 16. Since the goal of providing the second or new data storage system 16 is to generally provide enhanced or increased capabilities to the host or other data processing system 12, it is therefore desirable to migrate the data as quickly yet as unobtrusively

storage system 14 includes a predetermined number of drives or volumes, each drive or volume having a certain number of tracks or records, the second data storage system will be configured to imitate such a configuration.

Once the second data storage system 16 has determined that least one data element (such as a track) has not been copied from the old or first data storage system 14, the second data storage system 16 issues a request to the first data storage system 14 for the data element, step 206. Once received, the second data storage system 16 stores the data on the second data storage system 16 (typically in cache memory 18), step 208, updates the second data storage system data map/table 24, step 210, and returns to step 200 to determine whether or not there is a pending data read or write request from the host or other data processing system 12.

In one embodiment, the present invention contemplates that it may be desirable to "prefetch" data from the first data storage system 14 to the second data storage system 16. For example, the migrate or copy process 27 may, using commands native to the first data storage system 14, issue a prefetch or "sequential" data access request or command to the first data storage system 14, to cause the first data storage system 14 to continue to fetch or 'prefetch' a certain number of data elements to the cache memory 18 of the second data storage system 16. Such prefetching can significantly speed up the transfer of data between the first and second data storage systems 14,16 by greatly reducing the number of "read" commands which must be passed between the data storage systems.

In another embodiment, the migration process 27 may determine that one or more read requests from the host 12 are part of a sequence of such read requests. In such an instance, the channel process 27 may take the current address of data being requested by the host 12 and increase it by a predetermined number. For example, if the host 12 is currently requesting data from an address '411', the channel process 25 will issue a read request to the first data storage system 14 for the data at address 411. Generally

Claims

1. A system for providing on-line, transparent data migration between first and second data storage systems, comprising:

a first data storage device, holding a plurality of data elements; and

a second data storage device, coupled to said first data storage device, for independently migrating data from said first to said second data storage devices.

2. The system of claim 1 wherein said second data storage device is coupled to an external source of data, and responsive to said external source of data, for migrating data from said first to said second data storage devices.

3. The system of claim 1 wherein said second data storage device includes a data element map, said data element map for indicating whether at least one predetermined data element is stored on said second data storage device.

4. The system of claim 3 wherein said second data storage device is responsive to said indication from said data element map indicating whether said at least one predetermined data element is stored on said second data storage device, for selectively obtaining said at least one predetermined data element from said first data storage device independent of said external source.

5. The system of claim 4 wherein said second data storage device is responsive to one of a read or a write command accessing said at least one predetermined data element issued by said external source, for selectively obtaining at least said at least one predetermined data element from said first data storage device, and for storing said obtained data on said second data storage device.

from said data processing device directing said second data storage device to write a predetermined data element from said data processing device to a longer term data storage device, and responsive to an indication that said predetermined data element is not stored on said second data storage device, for issuing a data read command to said first data storage device for at least said predetermined data element.

11. The system of claim 10 wherein said second data storage device is responsive to an indication that said data write command received from said data processing device is for a predetermined data element which is a portion of a complete data element storage location, for issuing said data read command to said first data storage system.

12. The system of claim 11 wherein said second data storage device is responsive to an indication that said data write command received from said data processing device is for a predetermined data element which is an entire data element storage location, for writing said received data element to said second data storage device.

13. The system of claim 12 wherein said entire data element storage location includes a disk drive track.

14. The system of claim 6, further including a data storage device configuration device, coupled to said second data storage device, for providing data storage device configuration data to said second data storage device and to said second data storage device data element map.

15. The system of claim 6 wherein said data processing device includes at least one host computer.

16. The system of claim 6 wherein said data processing device includes a network.

17. A method for migrating data from a first data storage device to a second data storage device coupled to an external source of data and to said first data storage device, said first data storage device holding a plurality of data elements, at least some of said

searching said data element map by said second data storage device to determine if said at least one data element is stored on said second data storage device; and

responsive to said searching, selectively copying said at least one data element from said first to said second data storage device.

23. A method for migrating data from a first data storage device previously coupled to a data processing device, to a second data storage device presently coupled to said data processing device and to said first data storage device, the first data storage device including a plurality of data processing elements previously accessed by said data processing device, at least some of said plurality of data elements to be copied to said second data storage device, said second data storage device including a data element map including at least an indication of whether a data element having a predetermined data element storage location address is stored on said second data storage device, said method comprising the steps of:

configuring said second data storage device to include a plurality of data element storage location addresses corresponding to data element storage location addresses on said first data storage device;

receiving, by said second data storage device, from said data processing device, at least one of a data element read and a data element write request regarding at least one data element;

searching said data element map by said second data storage device to determine if said at least one data element is stored on said second data storage device; and

responsive to said searching, selectively copying said at least one data element from said first to said s second data storage device.

generating a data element read request to said first data storage device for said at least one data element requested to be written by said data processing device;

retrieving said at least one data element by said first data storage device, and providing said retrieved at least one data element to said second data storage device;

said second data storage device receiving said retrieved at least one data element from said first data storage device; and

said second data storage device writing said at least one data element to be written to said second data storage device.

26. The method of claim 23, wherein said first data storage device is previously coupled to said data processing device, and prior to coupling said second data storage device to said data processing device, performing the step of uncoupling said first data storage device from said data processing device.

27. The method of claim 23, wherein said step of receiving said requested at least one data element stored in said first data storage device by said second data storage device includes storing said received requested at least one data element in cache memory.

28. The method of claim 23, further including, after said step of receiving said requested at least one data element from said first data storage device, the step of updating said data element map.

29. The method of claim 23, further including, after the step of providing said requested at least one data element to said data processing device, the step of writing said requested at least one data element to said second data storage device.

30. The method of claim 23, further including, the steps of:

determining that said second data storage device is not completely busy at least responding to data

indication of whether a data element is stored on said second data storage device;

coupling said second data storage device to said data processing device;

coupling said second data storage device to said first data storage device;

configuring said second data storage device to include a plurality of data element storage location addresses corresponding to data element storage location addresses on said first data storage device;

determining that said second data storage device is not busy responding to data element requests from said data processing device;

reading said data element map of said second data storage device;

determining which data elements stored on said first data storage device are not stored on said second data storage device;

requesting said data elements which have not been copied to said second data storage device from said first data storage device;

storing said requested data elements on said second data storage device; and

updating said data element map to indicate that said stored data elements are stored on said second data storage device.

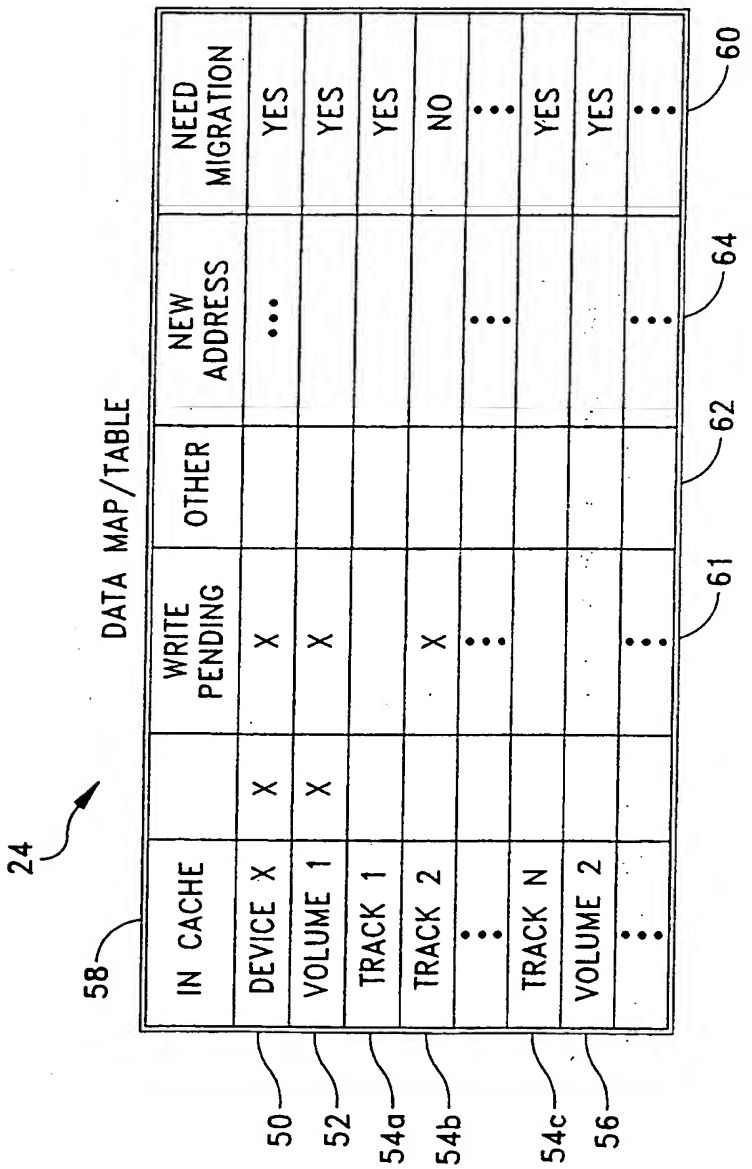

24
58
DATA MAP/TABLE
IN CACHE
WRITE PENDING
OTHER
NEW ADDRESS
NEED MIGRATION
DEVICE X
VOLUME 1
TRACK 1
TRACK 2
TRACK N
VOLUME 2
X
YES
NO
50
52
54a
54b
54c
56
60
61
62
64

FIG. 2

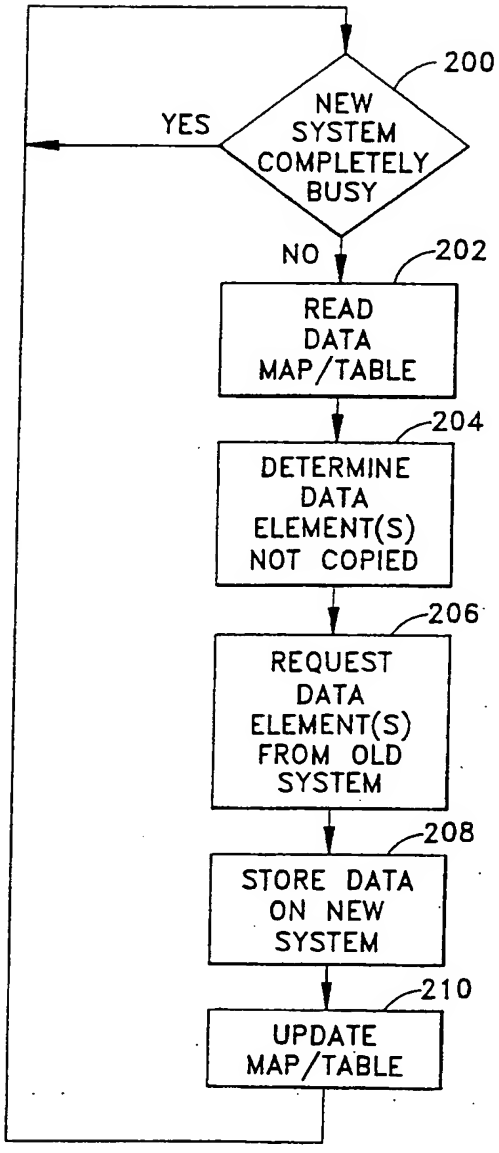

200
NEW SYSTEM COMPLETELY BUSY
YES
NO
202
READ DATA MAP/TABLE
204
DETERMINE DATA ELEMENT(S) NOT COPIED
206
REQUEST DATA ELEMENT(S) FROM OLD SYSTEM
208
STORE DATA ON NEW SYSTEM
210
UPDATE MAP/TABLE

FIG. 4